CA Journal of Cultural Analytics

Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora

Jo Guldi

12.20.18

Peer-Reviewed By: Ted Underwood & Anon.

Clusters: Theory

Article DOI: 10.22148/16.030

Dataverse DOI: 10.7910/DVN/BJNAPD

PDF DOI: 10.31235/osf.io/g286e

Journal ISSN: 2371-4549

Cite: Jo Guldi, "Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora," Journal of Cultural Analytics. *December 20, 2018.* DOI: 10.31235/osf.io/g286e

Whether they work with pages hand-selected from an archive, or keywords hand-selected from a digital corpus, scholars of all kinds wrestle with the issue of exemplarity.[1] Cherry-picked examples leave the scholar's conclusions vulnerable to charges of, at best, irrelevance, and at worst, malfeasance. In digital research, even seemingly simple queries are plagued with these pitfalls, which can only be addressed by critical thinking about the queries we propose, the algorithms we use, and the questions we ask of them. If digital research is approached responsibly, however, it allows the scholar to choose exemplary texts for reading with a precision and clarity unavailable to previous generations.

[1] The research documented in this article was graciously supported by the National Science Foundation IBSS Grant 1822091 and a fellowship from the Center of Creative Computation, Southern Methodist University.

Consider the case of searching for change in the language of property in the parliamentary debates. Displacement is a theme in modern Britain since the beginnings of the enclosure of traditional peasant commons in the middle ages, but in the nineteenth century, evictions and other displacements became more common, linked to an ideology of botanical improvement, the rise of utilitarian economics, the proliferation of race-based modalities of governance, and the creation of legal mechanisms to aid landlord-led improvement.[2] A dozen different secondary sources have located major changes in the understanding of property in nineteenth-century Britain, but no definitive method exists to synthesize these rival claims about when property changed and how.[3]

Looking for scholar-supplied keywords over time lends particular insights. The scholar, knowing the rising number and length of parliamentary speeches over the century, might choose to count keywords over time as a proportion of all words spoken each year.[4] The timeline shows an apparent eruption of debates about property rights after 1875, between the Landlord and Tenant Act (Ireland) of 1870 and the peasant insurrections known as the Irish Land War of the 1880s, followed by a further explosion of the terms after 1900. The importance of the yellow bar for "eviction" in the timeline (Figure 1) suggests that increasing mentions of tenants and landlords on the floor of parliament may have been driven by discussions of eviction at a time when cases of evicted peasants were being heard on the floor of parliament with great regularity.[5]

[2] Eric Stokes, *The English Utilitarians and India* (Oxford: Clarendon Press, 1959); Thomas M. Devine, *Clanship to Crofters' War* (Manchester: Manchester University Press, 1994); Fredrik Albritton Jonsson, *Enlightenment's Frontier* (New Haven: Yale University Press, 2013).

[3] Important texts include A. V. Dicey, "The Paradox of Land Law," *Law Quarterly Review* 21 (1905): 221-232; Avner Offer in *Property and Politics, 1870-1914* (Cambridge [Cambridgeshire]: Cambridge University Press, 1981); Paul Readman and Matthew Cragoe, eds., *The Land Question in Britain, 1750-1950* (Basingstoke, England; New York: Palgrave Macmillan, 2010); Paul Readman, *Land and Nation in England* (Woodbridge, UK: Boydell Press, 2008); David Steele, *Irish Land and British Politics* (London: Cambridge University Press, 1974); and L. Perry Curtis, *Depiction of Eviction in Ireland 1845-1910* (Dublin: University College Dublin Press, 2011).

[4] Ryan Vieira, *Time and Politics* (Oxford: Oxford University Press, 2015).

[5] Figures 1-3 were coded by Jo Guldi.

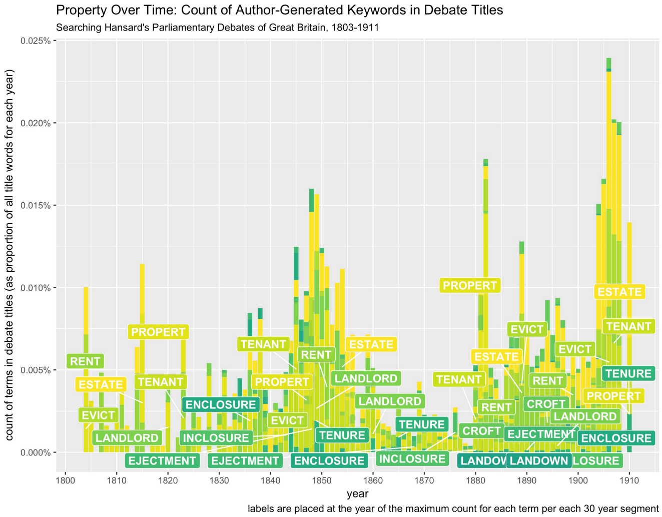


Figure 1.

There are problems with this search, however, for the visualization as a whole obscures conflations and exceptions. The keyword "tenant" pulls in not only those results about eviction, but also those about agricultural produce, property taxes, tithes, the franchise, and many other subjects only loosely related to relationships to property. Other general terms, for instance, "estate" or "landlord," similarly cast too wide a net, and the resulting list tells us very little about what changed when, how and why. Another omission is geographical and racial: the absence, from this list, of any words having to do with Indian or African tenure as opposed to Scottish, Irish, and English.

The scholar might try to enhance the results of Figure 1 by attempting to understand better the fate of individual words. Adding more information can refine this process towards a more reflective approach by, for example, using another scholarly apparatus—in this case the Oxford English Dictionary—to suggest the full variety of terms for property. In the revision, the search follows terms that include local relics of feudal culture (for instance "udal") and the lexicon of imperialism (for instance "zemindar"). From the OED keywords, the top ten most frequently-occurring terms will be plotted over time.

Revising the search process allows the scholar to investigate some additional questions of method and interpretation. Because the research question concerns displacement, the inquiry is unrelated to mentions of "rent" or "tenants" that pertain agricultural commodity prices, household taxation, or the franchise. One way to zero in on discussions of debates is to limit the inquiry to the ten debates in which each term is present the *most,* rather than every mention of landlords. Such a shift of perspective would have the benefit of examining the changing language in those debates that hinged the most on a lexicon of property.

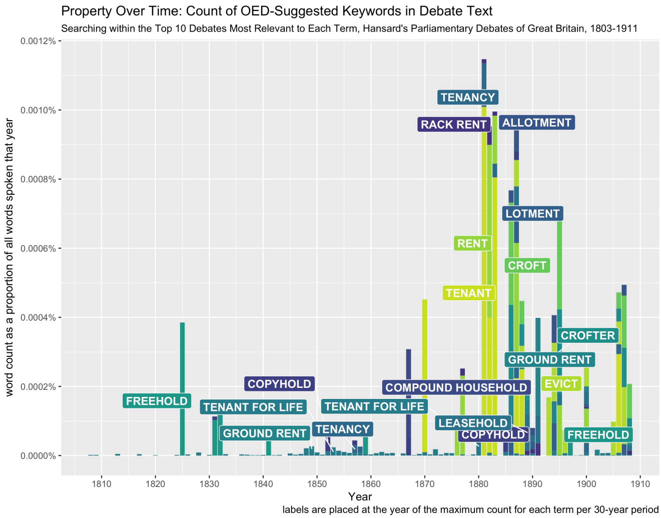


Figure 2.

The results of this second iteration (Figure 2) reinforce the sense of a transition from market-related terminology early in the century, where the language of "freehold" and "landhold" early in the century gave way to a different conversation about the "tenant" and "rent," mediated by the emergent concept of "ground rent." The later lexicon of property is more diverse, with a prominent discussion of the "crofter" and other tenant victims of "eviction," together with remedies such as "allotment," cresting between 1875 and 1890. The revised search reinforces the impression of an enormous explosion in mentions of tenant-landlord relationships around rent after the Irish Land Act of 1881), with a corresponding two-to-ten-fold increase in the usage of the new terms over the old, measured as a proportion of all words spoken in the period.

Like the first figure, however, Figure 2 is marked by the eurocentrism that typifies most debates in parliament, even in the age of empire. While the original list of keywords included "zemindar" and "ryot," counting the most frequently used words returned no Indian terms for landholding. Further approaches to the lexicon would be necessary to interpreting when and how any shift occurred in the lexicon of Indian, African, Pacific, or North American property.

From the viewpoint of a historian trying to recover the history of eviction in different times and places, the search still remains incomplete. The scholar could proceed by attempting another version of keyword search, perhaps selecting only the Indian terminology from the OED, or sampling only those debates that refer to India, Africa, Australia, or Canada by name in the title. Indeed, term-based search may be the wrong key to unlock the door of global eviction. Such a problem as this one requires the scholar to carefully consider the nature of the document base, to reckon with her choice of keywords, and even to consider other possible algorithms that might solve the riddle in a different way.

The foregoing discussion demonstrates both the necessity and the difficulty of critical thinking about digital searches. Neither the process of forming a question, nor the interpretation of that question, is automatic. Every word in a simple keyword search is open to unpacking; each group of words obscures others that might be elucidated by further searching, study, and reading.

Critical thinking about the words that supply a digital search lends strength and rigor to our research process. A process of critical engagement allows the scholar to correct for the proclivity to overinterpret a particular chart, that is, the tendency to construct a thesis from a single illustration of discontinuity. Iterative approaches and multiple tools are essential for controlling for the scholar's own subjectivity in encounters with the archive.

The two graphs in the introduction to this article demonstrate not a trajectory towards some ultimate treatment, but rather the ambling, iterative course of exploration that a scholar might take. Certain truths are revealed by one graph and other truths by another. Each illustration has limits which give way to new research questions, and the scholar must ultimately reconcile the findings of all the interventions to her original research project. The succession of charts forms a path through the data, as the scholar explores dimensions of the archive, explaining those findings to the reader. In the journey of critical search, the scholar engages with critical thinking at every step. The foregoing example also illustrates how multiple measures complement each other, enhancing the scholar's sense of *what* is being measured and *how* a particular search illuminates and disguises various dimensions of a canon.

This article calls for a critically-informed strategy for negotiating digital archives that is aligned with an understanding of how different algorithms determine particular perspectives on textual corpora from the past. Recently, for instance, Daniel Shore has shown how a dozen different algorithms produce a dozen different versions of the past.[6] Understanding digital algorithms as having this per-

[6] Daniel Shore, *Cyberformalism: Histories of Linguistic Forms in the Digital Archive* (Baltimore:

spectival ability to open up different dimensions of an archive reminds us that no search is complete until all of its aspects—the choice of keywords, the algorithm, the exceptions, and the particular texts taken as exemplary evidence of the result—have been subjected to iterative examination.

A call for critical thinking throughout the search works at odds with an empirical and scientific posture often taken by humanists who engage with digital tools, for instance in what we might call "proof of concept" articles in which a new tool is introduced to a new field. Such articles often stress the scientific correspondence between computerized generalizations and the reality of the archive, in order to validate a new method in the field.[7] In so doing, and stressing the "discovery" aspect of a new method, tools typically stress the unified nature of the reality produced by a particular archive and tool. It is unsurprising then that readers attached to suppressed voices from below might resist such tools, wondering indeed if they are instruments of a renewed imperialism of history by the pseudo-scientific fact.[8]

How digitally-enabled historians engage with macrohistory thus raises important issues about the interpretation of digital findings. Is the role of digital tools in "distant reading" necessarily to reveal a single, Apollonian, and definitive perspective on an archive or period of time? This article attempts to model a general process by which a scholar can approach the perspectival nature of algorithms. It argues for a critical, interpretive approach to digital tools based on iteration, where the scholar constantly uses the results of digital inquiry to investigate the question of what different options propose or produce. It asserts that each digital tool and the parameters with which is it is used provides its own perspectival approach on the vying lexicons, grammar, and ideas of the past. It urges a critical approach to the use of these digital tools, which is modeled in terms of three "macro-steps" in a process of engagement, choosing a seed text from the secondary literature, winnowing the results of the search, and guided reading in the results of the winnowing.

This article, therefore, proposes that the solution for better text-mining is not another algorithm, but a new attitude among scholars engaging with digital tech-

Johns Hopkins University Press, 2018).

[7] For instance, Kellen Funk and Lincoln A. Mullen, "The Spine of American Law: Digital Text Analysis and U.S. Legal Practice," *The American Historical Review* 123, no. 1 (February 1, 2018): 132-64; Matthew L. Jockers and David Mimno, "Significant Themes in 19th-Century Literature," *Poetics* 41, no. 6 (2013): 750-769; Lauren Klein and Jacob Eisenstein, "Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives," *Scholarly and Research Communication* 4, no. 3 (2013).

[8] For instance, Roopika Risam, "Beyond the Margins: Intersectionality and the Digital Humanities," *Digital Humanities Quarterly* 9:2 (2015).

niques, one that is modeled here under the phrase, "Critical Search." Critical search, like critical thinking, employs archives from a set of pre-existing social and political concerns, brokered through skepticism about the shifting meanings and hidden voices within any archive. As this article will argue, scholarship requires, above all, a careful process of iterative examination of the corpus, and iterative investigation, and the world of research by algorithm should follow these practices as well. No single text-mining technique—whether topic modeling, keyword searching, or statistical measures—is by itself sufficient for filling out social categories such as eviction, which was referred to under a variety of terms, many of them so general as to be eluding. Such an approach would fit entirely within a strain of historical method that sees historical agency as multiple and overlapping, and the task of historical interpretation as resurrecting and identifying the competing strains of agents vying for the past. Another perspective might assert that digital tools are generative and perspectival. They create visions and versions of the past that let us see other things.

Defined as an approach that incorporates critical thinking into the research process, Critical Search does not depend on a particular algorithm or set of algorithms, but rather suggests how questions of interpretation and scholarly selection permeate the entire process of applying digital tools and using their results. It argues, generally, for multiple and iterative engagement with different algorithms and tools for the purpose of generating a multiple and overlapping perspective on how different actors' lexicons, grammars, and discourses were changing across different periods of time.

The digitally-engaged process of Critical Search described in this article is designed to mirror a traditional history seminar, where students move from an assigned syllabus to a broader set of readings around a topic, followed by the identification of particular case studies for further reading. The student of history typically begins with some process of grouping together reading material based on her interests. Gradually, her reading moves wider, usually by using a variety of prostheses whose nature she understands well, including a card catalog (or now, more typically, its digitized version), or the research assistant tasked with assembling some primary sources for my next syllabus. If she does her job well, the resulting base of sources she uses is perfectly tailored to reflect her research question.[9] The social historian is thus faced with an enormous mass of

[9] The virtues of this expansive contextual reading, broad sifting of sources, and synthesis of evidence from different points of view have recently been articulated in the many-authored "Tuning Project" of the American Historical Association. See "AHA History Tuning Project: 2016 History Discipline Core,“ (accessed June 1, 2016). The critical search process proposed here, while crucial to the process of summarizing and dating events in history, is nevertheless fairly irrelevant to many fields in the humanities and information sciences. Critical search, as described here, would be alto-

possible records from which meaning could be extracted; hence the legitimacy of microhistorical approaches to social questions that follow individual lives and families as microcosms of larger dynamics of gender and class.[10]

Faced with imponderable archives, scholars need to use tools both to generalize and to narrow. Digital tools may help them to generalize about a corpus. Other tools may help them to identify particular texts that are symptomatic of larger trends, and to speak in specific ways about how a particular passage or set of words is exemplary of a larger whole. They may need to divide up massive archives into subcorpora, and to generalize within these smaller corpuses about the voices and trends they find there. For instance, historians who deal with official corpora need to be able to characterize what Robinson, Gallagher, and Denny called the "official mind" of the state, as well as to identify the texts recorded by the official mind, in the form of testimony, survey materials, and anthropological description, that reference social experience as it both resisted, remade, and was refashioned by the state.[11] Such tools as these would afford the digitally-aided scholar a set of advantageous techniques for the recovery and analysis of social experience through the mass-digitized archives so widely available today.

Humanistic research increasingly operates on collections where the scale of texts involved defies indexing by hand and results, increasingly, in the reliance upon technologies such as topic modeling as an intermediary between the researcher

gether unnecessary for a student of canonical politics who already knows the names of the actors who matter to him. Likewise, a student interrogating the female literary voice may only need to collect fifteen examples of novels by women to generalize her conclusions. The social historian, however, is responsible for portraying the range of voices related to a particular category, as well as adequately understanding the period for which her query is relevant, ideally by dating the first and peak expressions of her subject. For the literary scholar, mass extraction is irrelevant, and for the sociologist, broad winnowing of the scholarly record is unnecessary. For these reasons, the model of critical search proposed here differs from a more humanistic conception of research, for instance the one formulated by John Unsworth, where the choice and analysis of passages text from an already constrained sample – rather than the discovery of an appropriate subcorpus from an unreadable mass – is the critical factor under consideration. Unsworth describes a seven-fold list of unordered primitives, including"discovering," "annotating," "comparing," "referring," "sampling," "illustrating," and "representing," tasks that are suitable to a small collection such as the Blake Archive on which he was working at the time.

[10] An excellent recent example being Seth Koven, *The Matchgirl and the Heiress* (Princeton: Princeton University Press, 2016). The opposite approach, of course, also has validity: approaching the official record with the intent of extracting a case of how the assorting and abstracting mechanisms of modern government remade the life of the peasant. "Paradoxically, history from below may be (as mostly it has to be) achieved by examination from above," mused historian Peter Robb of his use of British state records to study the peasant of India. Peter G. Robb, *Ancient Rights and Future Comfort: Bihar, the Bengal Tenancy Act of 1885, and British Rule in India* (Richmond: Curzon, 1997), xxi.

[11] Ronald Robinson, John Gallagher, and Alice Denny, *Africa and the Victorians: The Climax of Imperialism* (Garden City, N.Y.: Anchor Books Doubleday, 1968).

and archival truth.[12] Scholars such as John Unsworth and Timothy Tangherlini have modelled the fit of the digital in the humanities by focus on the tasks of collecting and indexing.[13] As more researchers turn towards such technological intermediaries, certain agreements about the importance of critical awareness, the conventions of reviewing algorithm findings, and the documentation thereof become critical if researchers in the humanities are to function as a community.

Information retrieval is the subject of an extensive literature in library science, typically reduce the retrieval process to a single algorithm (for instance, tf-idf, the measure of terms relatively sparsely disseminated overall that are expressed in particular articles).[14] One of its conventions is the profiling of particular tools, one at a time. Journal articles about the digital humanities have frequently treated one digital toolkit at a time: consider how Lauren Klein introduced the topic model with the letters of Thomas Jefferson, or how Funk and Mullen explained the analysis of textual re-use in the American legal code.[15] In order to accept new knowledge provided by the abstraction of the of the topic model, readers need to first be persuaded that the abstraction of the topic model in some way provides new knowledge and that that knowledge can be verified in comparison with other, traditional means of learning about historical corpora. Most digital history articles to date, including some that I have written, conform to this extremely reductive convention of proving that a single tool is useful for understanding the past. But that convention need not dominate how we publish about the humanities, and it should not reduce our capacities to think in methodologically plural ways about the past.

In contrast to digital scholarship that profiles a single approach to the archive, this article will emphasize what happens as scholars move between questions, tools, texts, and provisional answers. This emphasis on the praxis of investigation represents an adjustment in approach from the strategy typically laid out in journal articles about digital history where a single tool is recommended for abstracting

[12] Gheorghe Muresan and David J. Harper, ”Topic Modeling for Mediated Access to Very Large Document Collections,” *Journal of the American Society for Information Science and Technology* 55, no. 10 (August 1, 2004): 892-910.

[13] John Unsworth, "Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?", in Symposium on "Humanities Computing: Formal methods, experimental practice" (King's College, London, May 13, 2000); Timothy R. Tangherlini, "The Folklore Macroscope: Challenges for a computational folkloristics," *Western Folklore*, 72(1) (2013): 7-27.

[14] Karen Spärck Jones, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," *Journal of Documentation* 28 (1972): 11-21; Stephen Robertson, "Understanding Inverse Document Frequency: On theoretical arguments for IDF," *Journal of Documentation* 60:5 (2004): 503-20.

[15] Klein and Eisenstein, ”Reading Thomas Jefferson with TopicViz“; Funk and Mullen,”The Spine of American Law”.

knowledge. To the degree to which we have a sense that it's appropriate to only apply one digital method at a time in analyzing digital corpora, that way of thinking reflects a convention that has arisen in journals around our conventions of explaining truth, a convention that is at odds with more sophisticated historical practice.

To interpret the range and depth of social experience requires an embedding of reductive techniques in a multiple-dimensional perspective on the past, which practice constitutes what I am calling critical search. What this means, in practice, is the integration of digital tools that *abstract* and *reduce* to a particular dimension—for example topic modeling with is abstraction of discourses of related words—with other tools that draw the reader's attention to still other dimensions of the text. A complementary and iterative approach thus provides an antidote to digital history as reductionism: when consulting topic models, it pays to use secondary sources to consider which texts to topic model and how to interpret the results; when using keyword search, it pays to use secondary sources to interpret the keyword search. Similarly, when topic modeling, it pays to also to reference keyword searches and keywords in context. Complementary and complicating abstractions and simplifications help the historian to convey the depth and breadth of their understanding of past events for the reader: and that has always been the challenge and promise of writing history.

Recently, certain scholars have pressed back against the reductionism of their methods, especially the assumption that scholars enter an archive already armed with an exhaustive understanding of the keywords, personal names, places, and dates that matter in a research project.[16] Indeed, recent standoffs over methods and theory have frequently taken the form of scholars insisting that new insight about historical change depends not merely upon the collection of new facts, but also upon insight into the changing history of human agency and institutions, realizations often opened up through engagement with critical theory.[17]

Critical search, as this article defines it, emerges from in between polarized positions of debate where some parties insist on critical thinking to the exclusion of new methods, and other voices focus strictly on the praxis of the method. One

[16] For instance, Nina Tahmasebi and others have investigated more generally the role of *a priori* knowledge in specifying the insights to be gained from the analysis of texts. Nina Tahmasebi et al., "Visions and Open Challenges for a Knowledge-Based Culturomics," *International Journal on Digital Libraries* 15, no. 2-4 (April 1, 2015): 169-87. There also exists a critique of this reductionism from within library science, for instance, Caleb Puckett, "Oh, the Humanities: Understanding information behavior to foster information literacy," *Emporia State Research Studies*, Vol. 46, no. 2 (2010): pp. 33-43.

[17] Ethan Kleinberg, Joan Wallach Scott, Gary Wilder, *Theses on Theory and History* (http://theoryrevolt.com, accessed June 1, 2018).

perspective is a reasoned appreciation of the power of algorithms in the humanistic research process. Another perspective is the critique of reductionism and the dangers of technological dependence given the multifarious nature of humanistic research. A third is the insistence upon the importance of critical perspective on agency and identity in the human past. From these three points of view emerges the necessity for critical search.

The Critical Search Process

In the process described in the rest of this article, an ideal set of stages alternating between algorithms, reading, and reflection formulate the humanist's research process. Critical search begins with critical reading of the sources, which then undergirds algorithmic modeling. Scholarly interpretation and computer-aided modeling alternate throughout the scholar's task. Modeling is then subjected to supervision, boundary-making, and guided reading, wherein a scholar then inspects the results of the algorithmic model for its accuracy as well as the vantages it opens on the material at hand. Statistical inspection of the results reveals the bias implicit in particular models. Documenting the state of the project in each of these categories opens the door to a truly transparent model of the scholarly project in a digital age.

The approach offered in this article—critical search—advocates that researchers proceed past preliminary reductionist models such as keyword searching and topic modeling, weaving together statistics, information theory, hermeneutics, critical theory, and critical reasoning into a model process, each step of which is open to inspection. The point of modeling a "critical search" is to offer reflections on a process of *narrowing* common to traditional research projects, which can guide the digital world in which researchers routinely need to constrain a large corpus around a particular question.

The categories of *Seeding, Winnowing,* and *Guided Reading* describe a sequence of research familiar to many professionals, who under the influence of method or theory constrain and broaden their reading on the basis of their findings. The resultant process typifies three general stages of opportunity for critical reflection on the search process and how the scholar has engaged with algorithms and primary and secondary sources—three places where those choices and can be usefully documented and described for other practitioners of history. In greater detail, the suggested categories are as follows:

Seeding.

The first question is what archive one addresses, and which known primary sources, dates, figures, or concepts govern the orientation to that archive. Considering these questions lends itself to a metaphor of inspecting and planting keywords, dates, and ideas passed on from elsewhere. The scholar's choices about which to engage necessarily change the shape of the later inquiry: patriarchal words will rarely reveal subaltern attitudes, for instance. In a digital process, the search process is generally also seeded with a choice of algorithm(s)—a topic model, keyword search, or statistical measurement of significance according to some abstraction. This process too will tend to shift the search in one direction or another, lumping or splitting the corpus according to some general mathematical or theoretical concept of discourse, lexicon, or cluster. Carefully documenting and discussing the choice of seeds—whether conceptual, semantic or algorithmic—represents a first opportunity for making transparent the search process.

Broad Winnowing.

Winnowing suggests work with the maturing fruit of a first round of searches, roughly working over the returns of some query to sort the wheat from the chaff, and discarding the less relevant options. In traditional research, the scholar chooses particular exemplary texts or characters for close reading only after engaging a wide variety of primary and secondary texts that allow her to map out how unusual they are; the researcher proceeds by working the source base to present other examples. In digital research, where an enormous corpus is generally present in every case, a researcher winnows with the algorithm, tuning and applying it to the results, testing how consistent are the results and how adequately they can be interpreted. In the process of working available algorithms and queries, the researcher will likely throw away many false positives or pieces of messy data, possibly trying the same search a dozen times with cleaner data and clearer results. Winnowing presents an opportunity for transparency in the documentation of how specific questions or algorithms work with a particular dataset and question.

Guided Reading.

Only at a mature stage of research does a researcher “harvest” the fruits of a research process as evidence of a shift over time, just as a gardener turns through tomatoes, discarding the ones too unripe, too moldy or bruised for consumption. In traditional research, a scholar inevitably discards or saves for later episodes from her work that are irrelevant to the research question as it becomes more and more targeted. In digital research, scholars must also choose which results bear not only upon the readership but also upon some historical question for her readership. The harvesting process in itself is laden with bias, and presents an opportunity for a scholar to explain in passing what was left out, and how much of the results as shown are the work of human sorting rather than the automatic detection-work of some algorithm.

Critical search thus humbly models the everyday interventions of traditional research and digital research, dividing them into the course of relatively natural seasons, each of which demands work, affords results, and offers an opportunity for transparent documentation of the choices made by scholars.

The process of critical search may be highly eclectic, and need not copy the steps laid out here: to engage in critical search is merely to insist on inquiring into the biases of different digital tools and their results at every step, constantly testing them and revising them with documentation. Later parts of this article will discuss, for example, an iterative research process that required successively reseeding, re-winnowing, and re-reading resulting samples of text from a corpus. To refer to portions of the model as stages or “seasons” is not intended to delimit or constrain. The resultant process may be either replicated simply (the way that by repeating a cookie recipe one procures cookies), repeated iteratively, or worked into the flows of inquiry that fork and take on new shapes with each pass. To divide them in three is merely to signal the many opportunities for documenting scholarly choice, and the biases that come with choice, as they are passed onto the next phases of work with data.

The bulk of this article explains, in greater detail, what the seasons of research look like in traditional and digital forms, and how algorithms and their the “fitting” to exploratory data analysis becomes part of the model. The article also follows the process of a critical search for texts about property in the parliamentary debates of nineteenth-century Britain as a case study. Provisional technical solutions form part of each of the three stages of critical search, but the algorithms presented here are deliberately chosen for their interpretability rather than as an assessment of the best algorithm from the continually evolving world of computer science.

Seeding with Words and Documents

Most scholars in some way start with names, dates, concepts, and words passed on from elsewhere, like those seeds that Indian peasant women sew into the hem of their garments for safe keeping from generation to generation. A strong component of scholarly research is likewise informed by tradition. Traditional scholars frequently renew a line of questions that previous generations posed before them, as when Peter Mandler opened one volume on reform with questions about the Constitution proposed by Oliver MacDonaugh.[18] Digital scholars, likewise, frequently consult earlier generations of Victorianists when determining which hand-picked keywords to follow for a quantitative study of Victorian moral ideals and how they changed.[19] Other scholars have started with contemporary documents, rather than words, using algorithmic matching to generate another set of documents linked by explicit textual re-use or similar thematic ideas.[20]

When a gardener goes to plant, she looks over the seeds and carefully chooses a few from a store. As those seeds begin to grow, she examines the hardier and weaker ones. Just so, the historian of Britain approaching Hansard's parliamentary debates may find herself pondering the categories that earlier generations of scholars have employed, and asking: Are the fundamental questions for text-mining parliament those of party, of individual personalities, of gender, race, or class, or of democracy or some other concept in general? Her choice is necessarily critically informed by changing theories in the discipline as well as her own temperament, politics, and interest.

The curation of the words and documents in any scholarly process tilt the results of the inquiry with the bias of the scholar and her world. From theory, from secondary readings, from prior knowledge of the canon, and from her own politics, the scholar approaches the vast unread with bias, that is, with certain ideas

[18] Peter Mandler, "Introduction," in Mandler, ed., *Liberty and Authority in Victorian Britain* (Oxford; New York: Oxford University Press, 2006).

[19] Frederick W. Gibbs and Daniel J. Cohen, "A Conversation with Data: Prospecting Victorian Words and Ideas," *Victorian Studies* 54, no. 1 (2011): 69-77; Bob Nicholson, ”Counting Culture; or, How to Read Victorian Newspapers from a Distance," *Journal of Victorian Culture* 17, no. 2 (June 1, 2012): 238-46; Thomas Lansdall-Welfare, Saatviga Sudhahar, Justin Lewis, James Thompson, and Nello Cristianini, ”Content Analysis of 150 Years of British Periodicals," *PNAS (Proceedings of the National Academy of the Sciences)* 114, no. 4 (January 9, 2017): E457-E465.

[20] Tangherlini and Leonard, ”Trawling in the Sea of the Great Unread”; Funk and Mullen, *”The Spine of American Law”*.

about which documents are the most important, which category of ideas, feelings, or names the best witness to change which tend to color the results. Those choices, in turn, govern all later computational output, and constrain the results of analysis in an important direction. The reality of bias is in no way different in traditional and digital search, for at every instance where the work of curation takes place, the scholar is generating an a priori reference point that shapes the search process.

Digital scholars have replicated this process of relating the "great unread" to a known canon of texts, figures, and events, sometimes by measuring the difference between canon and archive in aggregate, and sometimes by beginning with a familiar reference text and using algorithms to discover the most similar documents.[21] The latter approach implies, in Timothy Tangherlini's metaphor, that the researcher uses a canonical text as the "hook" to go "trawling" in the "ocean of the great unread."[22] In the proposed process, one known text becomes the basis for using other topic models as a finding aid by which to collect an assortment of other novels with similar content or themes.

In critical search, the biases that governed the choice of keywords, documents, and dates need to be made explicit and self-reflective as possible because they have such strong downstream effects. The work of bias in digital search may be examined by the process of "seeding" a search with particular keywords, where the scholar begins by counting keywords over time and then uses those counts to identify particular trends, documents, or passages for further reading.

A search seeded by keywords provides ample material for historical analysis, where a research question is already extremely narrow, for example, if we want to find every instance where newspapers mentioned Gladstone or to count conservative invocations of "manliness," as Luke Blaxill has done.[23] Critical thinking about Victorian gender and its presentation in parliament has provided a narrow set of seeds, appropriate to the time and place covered by the data, and the scholar has only to plant the seed and reap interesting results. In very specialized cases of this kind, further winnowing may not even be necessary before moving from keyword counts to further conclusions.

When scholars engage an abstact concept such as property or eviction, they may have a hard time distilling a broad discourse about property into a single keyword

[21] Mark Algee-Hewitt et al., *Canon/Archive: Large-Scale Dynamics in the Literary Field*, 2016.

[22] Timothy R. Tangherlini and Peter Leonard, "Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research," *Poetics*, Topic Models and the Cultural Sciences, 41, no. 6 (December 1, 2013): 725.

[23] Luke Blaxill, "Quantifying the Language of British Politics, 1880-1910," *Historical Research* 86, no. 232 (May 1, 2013): 313-41.

that measurably changed over time with interpretable results. The choice of any particular keyword tilts the results in particular directions. A scholar may, therefore, choose to begin by following a broader class of keywords, for instance "rent" and "tenure," but these approaches cause new problems. The keyword "tenant" includes not only debates about eviction, but also those about agricultural produce, property taxes, tithes, the franchise, and many other subjects only loosely related to relationships to property. Other general terms, for instance, "estate" or "landlord," similarly cast too wide a net, and the resulting list tells us very little about what changed when, how and why. Using these words as an index for debates draws attention to the wordiest debates, or the discussions characterized by their use of scholar-defined terminology, but not necessarily the ones in which the words' usage changed the most.

	debate	year	speechdate	wordcount
1	IMPROVEMENTS.	1883	1883-07-17	674
2	COMMITTEE. [FIFTH NIGHT.]	1881	1881-06-02	539
3	LAND TENURE BILL.	1906	1906-11-12	526
4	IRISH LAND LAW BILL [Lords]. [BILL 308.]	1887	1887-08-01	479
5	LAND PURCHASE (IRELAND) BILL.	1888	1888-11-27	479
6	IMPROVEMENTS,	1883	1883-07-19	462
7	SECOND READING. [FIRST NIGHT.]	1881	1881-08-01	452
8	ARREARS OF RENT (IRELAND) (recommitted) BILL. [BILL 213.]	1882	1882-07-11	415
9	SECOND READING. ADJOURNED DEBATE. [SECOND NIGHT.]	1870	1870-03-08	398
10	SECOND READING.	1870	1870-06-14	397

Table 1. Debates with highest wordcounts for "tenant" in Hansard, 1803-1908.

	debate	year	speechdate	wordcount
1	INDIA TENURE OF LAND IN MADRAS.	1854	1854-07-11	48
2	GOVERNMENT OF INDIA.	1853	1853-06-03	9
3	OBSERVATIONS.	1861	1861-05-31	7
4	TORTURE IN MADRAS.	1856	1856-04-14	6
5	GOVERNMENT OF INDIA ADJOURNED DEBATE.	1853	1853-06-06	4
6	REVENUES OF INDIA.	1856	1856-04-18	4
7	GOVERNMENT OF INDIA BILL ADJOURNED DEBATE (FOURTH NIGHT).	1853	1853-06-30	3
8	Madras Ryotwari System.	1902	1902-07-29	3
9	THE GOVERNMENT OF INDIA.	1853	1853-02-25	3
10	INDIAN FAMINE COMMISSION.	1902	1902-02-03	2

Table 2. Debates with highest wordcounts for "ryot" in Hansard, 1803-1908.

	debate	year	speechdate	wordcount
1	IMPROVEMENTS.	1883	1883-07-17	674
2	COMMITTEE. [FIFTH NIGHT.]	1881	1881-06-02	539
3	LAND TENURE BILL.	1906	1906-11-12	526
4	IRISH LAND LAW BILL [Lords]. [BILL 308.]	1887	1887-08-01	479
5	LAND PURCHASE (IRELAND) BILL.	1888	1888-11-27	479
6	IMPROVEMENTS,	1883	1883-07-19	462
7	SECOND READING. [FIRST NIGHT.]	1881	1881-08-01	452
8	ARREARS OF RENT (IRELAND) (recommitted) BILL. [BILL 213.]	1882	1882-07-11	415
9	SECOND READING. ADJOURNED DEBATE. [SECOND NIGHT.]	1870	1870-03-08	398
10	SECOND READING.	1870	1870-06-14	397

Table 3. Debates with highest wordcounts for "croft" in Hansard, 1803-1908.

"Tenant" also highlights a peculiarly Irish subject-matter, where the Irish Catholic tenants of English landlords became the subject of debate in the Land Laws of 1881 and 1887 and their later amendments. Careful seeding, therefore, requires at a minimum that the scholar use words that highlight some of the geographical diversity of the property issue in parliament, for example keyword searching for Scottish "crofts," and Indian "ryots" (Tables 1-3).[24] Even so, the scholar's bias towards well-documented territories like Scotland, India, and Ireland necessarily constrains the inquiry, with the native tenures of South Africa, New Zealand, and Jamaica nowhere in this list. The bias of the search terms still structures the results.

It becomes incumbent on the researcher to devise a way of moving from a larger list of terms for property to particular turning points. The researcher cannot move hastily from a list of keywords to the full collection of debates about property, then extracting place-names and proper nouns, for a reliable and exhaustive search of a particular corpus.

In some processes, it is not a word or a name that operates as a "seed," but rather a collection of documents whose lexicon will be "matched" by some algorithmic process. In this kind of work, the justification of the seed texts unambiguously colors the results that will later be returned, and the scholar must justify those choices in a discussion. In the case of studying texts about property and eviction in Hansard's parliamentary debates of the UK, four reports relatively well-cited in the secondary literature were chosen as a "seed." They were chosen to eliminate bias, given relatively geographic representation - two reports for Ireland, one for Scotland, and one for England—and chronological representation—one from the 1840s, and three from the 1880s. A scholar writing about the results of the algorithm in question cannot present her work as final: "seeding" the process with slightly different texts or excerpts from those documents would result in totally different findings.

A seed can even go beyond a definite keyword or text, to be an evolving category. For Tim Tangherlini, the seed was known works of fiction in Danish literature—a static category—and topic models were used to harvest less-known works on the same subject. For Simon DeDeo and Rebecca Spang, the seed was "the future" relative to each year in the *debats* of the French Revolution—an evolving category—and statistical divergence was used as a measure of how much any speaker anticipated what the French government would be talking about in months and years to come. The choice of seed documents—whether canonical reports, or

[24] Tables 1-3 were coded by Jo Guldi.

futurity itself—governs intimately the analysis that follows.

The point of identifying seeding as a particular phase in the process is to call scholars' attention to the need to document and justify the determinative choices made at the beginning of research. The goal is not to eliminate bias altogether, but to raise the reader's critical awareness of the exclusions implied by certain choices of words or documents, thus opening up later inquiry by others about how the algorithms might have been programmed to return different results.

A critical search with keywords or documents moves from naïvely proposing a word or document and the "answer" to an algorithmic search, to a rich description of why *these words* or *that document* will illuminate further inspection of the past. The process remains potentially fraught and given to further argumentation and inquiry at every stage, and the problem of interpretation never disappears.

The documentation of these choices is not standard practice in history or in the humanities in general. Neither in traditional research or digital research is it common to reference the card catalogue used, the key terms searched for, the dead ends taken, and the routes that were most profitable along the way. Traditional scholars may have felt that their reading habits were more important. In digital scholarship, however, documenting the choice of seeds is crucial to making a query reproducible.

"Reproducibility" of a query may be a new virtue for some humanists, especially for those who work in the realm of interpretation, polysemy, and affect; but even those communities too should appreciate the role that footnotes have long played in documenting conversations and allowing communities of consensus to emerge. In the discipline of History at least, the reproducibility of individual visits to the archive has long been a standard of truth-making. Digital scholars, who regularly share both data and code with others, have an opportunity to convert that standard of truthful, replicable analysis by demonstrating, as they lay out their analysis, that the choices of seed texts and algorithms correspond with both the ideas at stake in their analysis and with the analysis actually performed by their code. Radical transparency about the choices upon which our arguments hinge promises to solidify nothing less than the humanities' role as defenders of mutually-agreed-upon truth—and of multiple possible avenues towards that truth.

In any event, making the invisible choices behind an analysis visible also offers a pedagogical opportunity for authors to educate their readers about how skillful scholarship is done. Transparency also makes the choices of scholarly analysis radically accessible to the classroom, where students of code should be instructed to try out, for themselves, alterative choices of seed texts and algorithms, so as to better understand the consequences of particular inquiries. For all of these reasons, documenting the choice of seed words, texts, and queries opens up the route for mere search to become critical search.

Seeding with Algorithms

The digital scholar typically uses the computational “match” in the way that earlier generations of historians used a card catalog or a bibliography or book indices: all of these technologies are tools for more accurately discerning an overview of available materials and also finding particular texts that will later form the basis of an argument.

In both traditional and digital research, the process of searching can, in theory, go on forever. Choices have to be made; the process is constrained. One tries a topic model and a keyword search, but no scholar tries out every possible algorithm or variation thereupon.

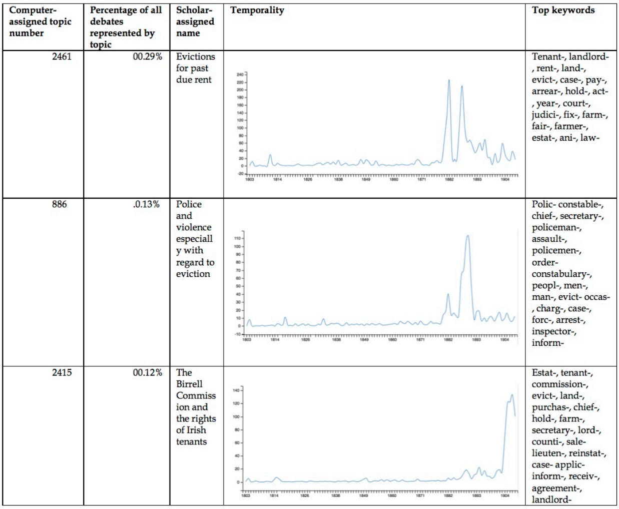
Trying out more than one approach opens the door to a critical perspective on how each algorithm, or each setting on an algorithm, tilts the results of research in a different direction. The possibility of reading through different prisms allows the scholar to begin to draw conclusions about the corpus that are based, inherently, on the breadth of sampling texts from the entire range of debate.

“Seeding” the process with other kinds of algorithms can illuminate different dimensions of the data, although it cannot escape entirely from reductionism. For instance, the technique of topic modeling has been routinely used to index and analyze digitized textual corpora, from Thomas Jefferson’s letters, to American newspapers, to ads for runaway slaves.[25] Topic models identify semantic similarities in collections of words that are used together, and they can even identify words that are used in multiple senses. For this reason, topic models are ideal for dating overlapping, competing discourses that use many of the same terms in slightly different ways. In a 500-topic model of the Hansard debates, eviction shows up among the top keywords for three topics relating to Ireland (Table

[25] David J. Newman and Sharon Block, “Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper,” *Journal of the American Society for Information Science and Technology* 57, no. 6 (April 1, 2006): 753-67; Lauren Klein and Jacob Eisenstein, ”Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives,” *Scholarly and Research Communication* 4, no. 3 (2013); Cameron Blevins, ”Space, Nation, and the Triumph of Region: A View of the World from Houston,” *Journal of American History* 101, no. 1 (June 1, 2014): 122-47.

4),[26] one linked to the disputes over tenants' rights to compensation in the 1850s, another linked to evictions for arrears of rent during the "Land War" of the 1880s, and a third related to the Estates Purchase Act of 1903 and the Estates Commissioners who heard previously evicted tenants' claims to the right of reinstatement. The same abstract keywords—"tenant," "estate," "property," and "landlord"—are employed in slightly different senses in each topic. The topic model thus picks up patterns that would be invisible to the scholar armed only with keyword search.

The power of viewing those words through the topic model is the computer's ability to pull apart slightly different discourses into regular patterns of keyword co-occurrence, each of which has its own chronology. Comparing different topics to each other suggests an evolution of the changing focus of discussions about land reform in Ireland after 1880, as it moved from the evicted tenants themselves, to the use of police and their clashes with tenants delinquent in their rents, to the eventual resolution in the form of state-led buyouts that provisioned former tenants with farms. Topic model thus becomes an aid to discerning discourse rather than keyword usage, and topic models once compared allow the scholar some insight into the life-cycle of long-term competing discourses.



Computer-assigned topic number	Percentage of all debates represented by topic	Scholar-assigned name	Temporality	Top keywords
2461	00.29%	Evictions for past due rent		Tenant-, landlord-, rent-, land-, evict-, case-, pay-, arrear-, hold-, act-, year-, court-, judici-, fix-, farm-, fair-, farmer-, estat-, ani-, law-
886	.0.13%	Police and violence especially with regard to eviction		Polic- constable-, chief-, secretary-, policeman-, assault-, policemen-, order- constabulary-, peopl-, men-, man-, evict- occas-, charg-, case-, forc-, arrest-, inspector-, inform-
2415	00.12%	The Birrell Commission and the rights of Irish tenants		Estat-, tenant-, commission-, evict-, land-, purchas-, chief-, hold-, farm-, secretary-, lord-, counti-, sale- lieuten-, reinstat-, case- applic- inform-, receiv-, agreement-, landlord-

Table 4. Eviction topics from a 500-topic model of Hansard, 1803-1908.

Even while the topic model illuminates some aspects of a historical transition,

[26] Table 4 was developed by Jo Guldi on the basis of an interface developed by Poom Chiarawongse, undergraduate, Brown University.

however, it simultaneously masks others. As with keyword search, topic models in this case have obscured the geographical diversity of British empire in the debates. A fine-grained topic model of all Hansard at 500 topics evidences discourses about the British occupation of Ireland, and not in Scotland's Highland Clearances, England's enclosure of the peasant commons, and India's railroad building, but nothing, once more, about Jamaican former slaves and their small farms, or African native tenure. Topic models can produce simplicity by indexing discourses over time, but the simplicity of reduction can be both a strength and a weakness.

The topic model, like the keyword search, becomes most useful in the course of synthesis, when a scholar wants to reduce a research question to a single typical episode of history that may be followed more deeply. When investing in an archive, scholars consult secondary sources, critical thinking, and interpretation; digital scholars need to do as much as well, moving from topic modeling to other dimensions of textual reduction, including other secondary sources and the keyword search. Keyword search and topic model enhance each other by representing separate aspects of a digital corpus: where a topic model abstracts a set of discourses whose representation changes over time, the keyword search, helps the user of the topic model to understand the particular life story of each individual keyword within the topic model. The scholar who wants to understand one dimension of a text is aided by moving back and forth between each abstraction.

The choice of method used to plant seed texts creates biases that will reverberate through the rest of the search process, for instance, modeling the domain of texts as a galaxy in which particular systems of affinity emerge, or as a yardstick with two poles. The booming search engine industry has funded a flourishing tide of studies about matching text to similar texts in machine learning, and here scholars have many tools that they can adopt, from packages for automatically "matching" documents based on a black-box similarity ranking, to software packages that provide similar matches for the technically adept.

In galaxy-type analysis, affinities between texts are represented as clusters, which can be traced by tools such as k-means clustering or topic modeling, where the algorithm has been designed to create probable clusters of documents that mirror human discourses, some large and some small, a principle that typically works for those interested in the discursive nature of a corpus.[27] Humanistic scholars have identified topic modeling with the scholarly reading of "discourses," and defended its logic as compatible with scholarly projects in the humanities.[28] Tools

[27] Tangherlini and Leonard, "Trawling in the Sea of the Great Unread"; Roe et al.,"Discourses and Disciplines in the Enlightenment".

[28] Roe and Gladstone winnow the rhetorical from substantive by working only with the nouns in

that operate on the level of figures of speech, for example, collocation analysis, may be more useful for identifying a *rhetorical* match that finds similar ways of speaking can be differentiated from the *substantive* match about particular subject fields.[29] Another set of tools, classified as "word embeddings," promise to transcend both categories, but have been less routinely studied in the digital humanities.[30] The weakness of all cluster-type analyses is—as with topic modeling—the reduction of texts into generalizations. To counteract the limits of clustering, a scholar can use a yardstick-type algorithm that simplifies the corpus according to a fundamentally different metaphor of abstraction.

Yardstick-type analysis include measures from information theory that impose a spectrum of order onto a corpus, arranging documents according to their linear proximity to some pole represented by another body of text.[31] In many fields, divergence measures have served as a fundamental metric of difference where difference is comprised of many factors whose expression is hard to describe. Divergence measures treat any two texts as a distribution of probabilities and arrive at an artificial *number* representing the distance, based on similar expression of the lexicon as a whole. The flexibility of creating a metric where none previously existed affords the making of structural comparisons in domains where comparison was hitherto available solely on a qualitative basis.

The scholar's choice of algorithm colors the results by revealing different dimensions of the experience of reading, whether topic-driven affinity or divergence-driven polarity. Where a topic model will provide a clustered relationship characterized by affinity groups, using divergence produces a more polarized answer, i.e. the "most close" texts to the seed texts are those that linearly ranked on a spectrum of distance according to how much they literally share the same word-

the *Éncyclopedie.* Would this strategy work as well in a legal context, when verbs and adverbs govern a field of procedures? Glenn Roe, Clovis Gladstone, and Robert Morrissey, "Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie," *Frontiers in Digital Humanities* 2 (2016).

[29] Alix Rule, Jean-Philippe Cointet, and Peter S. Bearman, "Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790-2014," *PNAS* 112:35 (September 1, 2015): 10837-10844.

[30] Ryan Heuser, "Word Vectors in the Eighteenth Century, Episode 1: Concepts." *Adventures of the Virtual* (14 Apr 2016), and "Word Vectors in the Eighteenth Century, Episode 2: Methods," *Adventures of the Virtual* (1 Jun 2016); Andrey Borisovich Kutuzov, Elizaveta Kuzmenko, and Anna Marakasova, "Exploration of Register-Dependent Lexical Semantics Using Word Embeddings," in *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 2016, 26-34.

[31] Alexander T. J. Barron et al., "Individuals, Institutions, and Innovation in the Debates of the French Revolution," *Proceedings of the National Academy of Sciences* 115, no. 18 (May 1, 2018): 4607-12; Sara Klingenstein, Tim Hitchcock, and Simon DeDeo, "The Civilizing Process in London's Old Bailey," *Proceedings of the National Academy of Sciences* 111, no. 26 (July 1, 2014): 9419-24.

usages. Divergence is the more transparent of the two, as it allows the scholarly to directly explain the relationship between two "similar" texts and the words they share in common.

Whichever approach the scholar chooses, the choice of algorithm is likely to produce entirely different results. Careful comparison of seed algorithms ideally enhances the scholar's sensitivity to different possible dimensions of the archive, for instance, the synthesis of discourses by topic model and the comparison of changing lexicons by divergence. Selecting different algorithms, if done critically and carefully, will direct the scholar to an archival base that does not merely reduplicate the language or concerns of other scholars. To become aware of the particular dimensions of each algorithm is to raise the possibility of documenting for other scholars a critical dimension of scholarship in terms of the particular choices and biases at play and how they can be reworked around different questions.

Each part of this process is potentially open to debate. The choice of keyword search, topic modeling, and divergence measure is to some extent arbitrary, but these three were selected as covering a wide ground of supervised and unsupervised work, and "galaxy" and "yardstick" type of measures.

In the case study of searching for eviction in parliament, both keyword search, topic modeling, and divergence measurement were used for a first pass through the Hansard corpus. Keyword search and topic modeling provided different, complementary answers to when and how the discourse of property changed over time. Divergence measurement resulted in a potential subcorpus of parliamentary texts that were judged by the computer to be close in lexicon to known primary sources about property, and this subcorpus was reserved for inspection later.

Within each algorithm, the scholar selects particular settings, and these too are open to inspection: there's the question of which scholarly terms are searched as keywords; of how many topics the algorithm is asked to return; of how the text is stemmed and stop-worded. If divergence is used, the scholar must decide on a mathematical formula for measuring similarity, whether the object of measurement will be the lexicon as opposed to ngram, skip-gram, or topic; and whether all words will be used or a particular lexicon with general vocabulary stop-worded out.[32]

In the case of the property question, measurements of similarity were taken on the basis of how common each parliamentary debate was based on the measure-

[32] The *philentropy* package in R collects at least twenty algorithms that could be used with different results). See Hajk-Georg Drost, "Introduction," accessed November 5, 2018.

ment of the probability expression of individual words. Measuring the distance between documents in terms of words, it was hoped, would capture a shifting vocabulary of property, whereas multiple-length n-grams might make a more appropriate measurement for rhetorical similarity. Place names and personal names were stop-worded out to prevent the returns from reduplicating the Irish-Scottish geography of the seed texts. Individual debates were compared, as opposed to individual speeches or paragraphs, in the hopes of finding subjects of debate that matched the question of property. Three common mathematical formulae were used that matched those commonly used by other scholars of text mining.

Selecting the seeds is followed by another stage in the process that entails the choice of algorithm for finding patterns in the broader corpus. The point of the next step is to begin the process of winnowing the full corpus of documents to a smaller subcorpus.

Broad Winnowing

Winnowing is an agricultural metaphor for sorting what is valuable to the scholar from what is not. To highlight it as part of a process of critical search is to underscore the fact that in some processes, the results of pattern-recognition are more useful than others, and any choice entails adding layers of bias. Several kinds of winnowing may be required, including adjusting the algorithm and its mathematics to reveal the biases implicit in one assessment over others, and down sampling the results of an algorithmic search. In either case, critical assessment of the winnowing process requires the scholar to explain the choices made in analyzing, preferably taking some measures to document how much the bias was tilted by a particular set of choices.

In this form of scholarship, winnowing typically takes the form of a discursive encounter that justifies the scholar's attention to certain objects in the archive. Traditional scholars make critical choices about their work by engaging in critical theory, sometimes more explicitly than others; in some cases, for instance, historians may depart from a critical reading based on social theory, but at other times, concern with theories of gender, race, and class may inform their approach to historiography and drive them to ask new questions. On the basis of this orientation, they may seek out particular lost voices in the archive, or trace particular encounters. Inn traditional research, winnowing is driven by critical theory and shaped by scholarly attention.

Winnowing with digital tools, by contrast, begins as a technical question: what adjustments can be made to this algorithm to “fit” my question or my data? What bias does each adjustment confer onto the results of my research? This technical question may be informed by an engagement with critical theory or political ethics, as for the traditional scholar. Indeed, in the case study of a search for eviction in parliament, the original engagement with eviction has its roots in a question about how policy reacted to an era of displacement, and whether the experiences of dispossession reached the corridors of power. That critical question, in turn, drives a technical problem of identifying the legislative debates that concern property and dispossession. In digital research, winnowing is motivated by critical theory and guided by the scholar’s skill at matching research questions with the tools of information retrieval.

The scholar may proceed by comparing the results of different algorithms, or by adjusting the settings of particular algorithms, for instance, trying out different divergence measures to get a sense of how varied their results might be. In no case it is clear that there is an objectively *right* setting—a single right metric or right tool, an objectively “right” granularity of topic modeling. Rather, by thrashing the data with different tools, the digital scholar obtains insight into the bias of the tools themselves, and the variety of answers they can produce. Because algorithms are rarely a perfect fit for scholarly questions, the scholar may adjust algorithms repeatedly in the process of honing in on a particular question, all the while learning more about the “fit” of a particular tool. At the heart of this matter is choosing a method that is tight enough that the results usefully answer a scholarly question, while loose enough that the scholar may potentially be made aware of unknown discoveries. These issues bring us immediately to the next two categories.

Winnowing may entail comparing variant settings on a particular algorithm, for example, the precise mathematical formula used to calculate similarity in a divergence measure, so that the scholar understands the bias related to using particular mathematical choices. In the case study of a search for eviction in parliament, three common divergence algorithms were used to measure lexicon similarity between documents. In order to explore the dimensions of chronology opened up by each choice of measurement, the results of the three mathematical formulae were compared. The results of divergence measure also depend upon the user’s choice of constraints for boundedness, so different cut-offs of similarity were examined in order to understand some of the different interpretations that might result.

Dedicating a phase of research to studying the effects of the first pass of broad winnowing gives the researcher an important opportunity to specify, critique, and interpret the relationship between the seed texts and the full corpus regarding the raw difference encoded by some algorithm. In the course of this iterative narrowing, the scholar has repeated opportunities to study the consequences of each pass with the algorithm. In the case study about property in Hansard, divergence between the four reports used as seed texts and Hansard as a whole can be represented as a histogram of speeches each of which has a different distance from the seed text. The resulting distribution groups the most similar debates to the left of the x-axis, and the least similar to the right of the x-axis (Figure 3).[33] The y-axis tells us how many debates are in each category: a few are very similar to the seed texts, many are somewhere in between, and a few are very far away. The KL measure classifies three out of the four seed documents as "more similar" to the rest of Hansard compared with the Bessborough Report, shown in yellow in Figure 3.

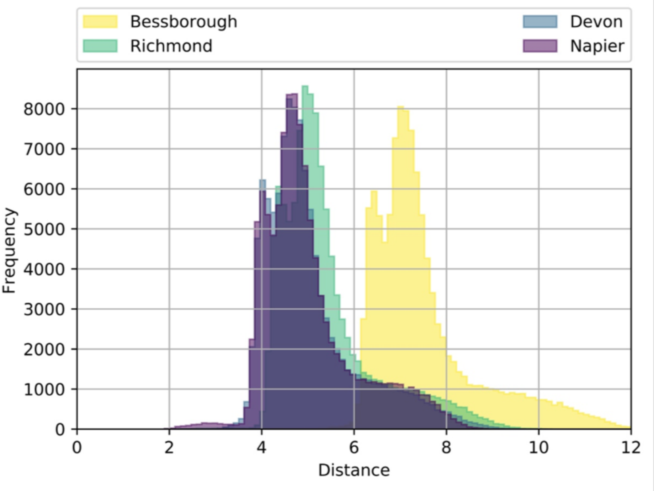


Figure 3. Kullback-Leibler divergence between Hansard debates as a whole and each seed report (distinguished by color).

As soon as a distribution of similarity has been created, the scholar obtains new information about how a particular metric describes the corpus as a whole. This information is critical for transparently presenting the choice of measure.

[33] Figures 3-4 were coded by Ashley Lee of Brown Data Science.

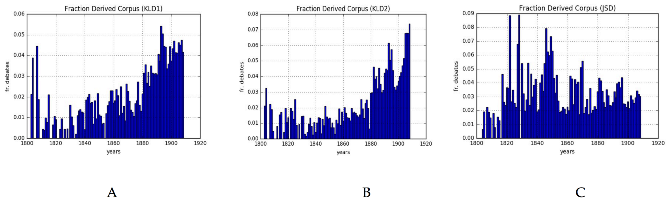


Figure 4. When did parliament debate property during the nineteenth century? Three different measures give three different answers.

A blue histogram gives the number of parliamentary debates classified as the 1% most similar to the Seed Corpus. The x-axis gives the year (1800-1910); the y-axis gives the fraction of all debates that were classed as 1% most similar to the Seed Corpus within that year. The three measures applied are: (A) Kullback-Leibler Divergence, (B) Symmetrical Kullback-Leibler Divergence, and (C) Jenson-Shannon Divergence.

Chronological diagrams help the researcher to understand how choosing a particular measure of similarity may influence the resulting chronology of documents returned. While mathematically similar, different measures of divergence will produce an utterly different sample of historical texts. For instance, compare how three different divergence measures answer the question, "When was property debated by parliament?" (Figure 4). In each case, the histogram shows how many debates are classified as being amongst the debates that are the 1% most similar in their distribution of language when compared to the seed corpus. Three different measures of "similarity" produce three utterly different chronologies of when property was debated. For instance, two formulae for measuring similarity returned results weighted towards the end of the century (Figure 4: a-b), while another formula returned results marked by four outbursts over the same period (Figure 4: c). In some cases, comparing these results will be sufficient to tell the scholar which measure "fits" their question. Broadly winnowing offers an opportunity for the scholar to "check" that the computer's first pass at the corpus is reasonable.

Diagrams that show the distribution of similarity the chronology of the most similar texts can be used along with topic models, top words, and other reductive synopses to judge different algorithms with regard to their "fit" for the question at hand. The chief problem satisfied by distributions and chronological histograms is to raise the scholar's consciousness of a *general bias,* reasonable or unreasonable, associated with unsupervised research associated with one particular similarity measure.

In approaching a problem through critical search, it is the scholar's duty to document those choices and the inevitable bias thereby entailed. Highlighting the bias of different algorithms does not necessarily resolve scholarly questions: it may leave them unanswered for later work. In the case of the search for "property" in *Hansard,* the results of modeling and abstracting the returns from three different divergence measures were far from conclusive. Distributions and chronological diagrams offered little guidance about which mathematical measure was ultimately better suited to the question of distance. The abstraction of mathematical clustering and measuring is simply too far from the scholar's *techne* of critical reading and discourse identification for a best "fit" to be determined by yet another abstraction of the same corpus. Winnowing the mathematics of the algorithm, in this case, may require the scholar to choose a particular approach to the corpus without being able to defend it, while making clear that the resulting chronology is definitively biased thereby.

None of these choices add up to a single model process; the choices made for a case study of property might not work for another inquiry or another archive. In arguing for critical search, the point is not to define an ideal set of algorithms or approaches, but to illustrate how each metric of measurement opens up a potential new window for the interpretation of the past. Whatever the choice of algorithm - whether divergence or another measure of similarity - the approach to the algorithm is crucial. In fact, any classification of texts—including topic modeling and word collocation—could be interpreted in a similar light, as a preliminary binary division of texts that could be refined, over time, through repeated passes through the process of critical search proposed here. Again, the responsibility of the scholar in critical search is that of documentation.

Guided Reading

Just as the gardener picks over moldy and damaged fruit for those good for eating and those good for pie, so too the scholar chooses the quotations with the clearest bearing on her questions. Digital scholars too must reckon with the choice of which findings to present. At this stage in the process, the scholar carefully inspects the results returned by a search process, sometimes sampling them, sometimes generalizing about them (for instance by counting keywords again or topic modeling).

The point of calling the last stage "guided reading" is, of course, to underscore that the end of critical search is actual *reading* of particular texts, including close examination of particular episodes, characters, turns of phrase, or tensions in the original primary-source documents. The visualizations and timelines produced at earlier points in the analysis may provide context, but they usually do not offer the end goal: an understanding of history. For most scholars, that understanding will only feel complete once they have both a macroscopic overview of change as well as some understanding of the individual lives and struggles of some micro-historical encounter.

The guided reading stage of critical search thus refers to the process of moving from algorithms to particular texts, followed by reading and interpreting them with the skills of a traditional researcher, that is, with critical thinking. The researcher uses the contextual overview of some algorithm to identify particular documents for inspection. A minimal exercise in guided reading would be to follow a keyword search to the year when a particular term took off, that is, the first year that saw exponential growth in the appearance of that term. Within that year, the researcher might locate the three documents of the corpus where those keywords form the highest proportion of words. She would read with confidence that these documents were likely candidates for having influenced others in their use of the term.

The promise of guided reading, which follows algorithmic contextualization, is the detection of documents that bear a significant relationship to a question. Based on the foregoing exercise in contextualization, the researcher can encounter the documents with confidence that they offer a potentially meaningful expression of the words in question.

At each stage of sampling and reading, the researcher gained insight into the promise and shortcomings of different algorithms. At the same time, she also gathered and read documents that the algorithm has classified as possible versions of an exemplary answer to her research question.

Through sampled reading, she gained sensitivity into the usefulness of working with the algorithm's thresholds, which indeed classified documents more closely related to known primary sources and more surprising. In the course of this research process, the researcher came to understand that some level of surprise in the algorithm (the 20-25% threshold of similarity) produced texts that were more instructive for guided reading. Iterative encounters with the algorithm and reading allowed the researcher to find documents that fit best with her questions.

Resampling, or Iteration of the Process

At the end of the second world war, the statistician George Box argued that "iteration" at the heart of statistical inquiry into science. Box set forward a general theory of critical reasoning, investigated by way of Francis Bacon and Ovid, in which a crucial component of good science was the ability to take measures to tailor a process such that new discoveries were possible. In Box's conceit, Pygmalion the scientist must not fall in love with his model. Resampling, in Box's view, was the psychological prophylactic that would prevent the researcher from entering into a "feedback loop" where the results were predetermined from the beginning of the experiment. Iterative resampling of the data at different levels allowed the scientist to judge the full breadth of the data. Each resampling would allow the researcher to constrain the experiment to eliminate error, but only in such minimal ways as to keep the experiment consistently open to unforeseen results.[34]

Across the disciplines, recent scholarship has insisted on the importance of systematic iteration where data-driven questions are at stake. Sociologists James Evans and Pablo Aceves more recently have emphasized the importance in social science of iterating between "theory confirmation" and "theory discovery."[35] Literary scholar Richard Jean So has also persuasively argued for the importance of learning about iteration in data-driven processes in the digital humanities.[36] Documented iteration has become increasingly crucial to the research process as humanities and social science scholars engage algorithms and quantitative thinking.

Critical search in itself attunes the scholar's sensitivity to the bias and perspectival nature of particular algorithms. In many cases, however, one pass through the algorithms is not enough. Keyword search, topic models, and divergence measures may all be used to narrow a corpus down to a smaller body of texts, for example identifying a particular decade of interest. In order to precisely "tune" the algorithms to the researcher's question, successive rounds of the critical search process may be necessary.

Iterative seeding and winnowing provides safety barrier against naïvely embracing the results of computational algorithm. At present, it is unclear how dependable most of our best tools for modeling text are, and where careful limits need to

[34] George E. P. Box, "Science and Statistics," *Journal of the American Statistical Association* 71, no. 356 (1976): 791-799.

[35] James A. Evans and Pedro Aceves, "Machine Translation: Mining Text for Social Theory," *Annual Review of Sociology* 42, no. 1 (2016): 21-50, 29.

[36] Richard Jean So, " 'All Models are Wrong,' " *PMLA* 132.3 (2017), 668-673.

be provided. For instance, computer scientists who deal with topic models have themselves called for more studies of whether, why, and how the topic model aligns with insights gained in traditional approaches. Eric Baumer and his colleagues have warned that there is "little reason to expect that the word distributions in topic models would align in any meaningful way with human interpretations."[37] Iterative winnowing and reading offer insurance against embracing foolhardy conclusions from digital processes. A truly critical search requires human supervision wherever the fit between algorithms and humanistic questions is unclear.

Seeding, winnowing and guided reading together may take the form of iterative encounters with the results of an algorithmic search. The researcher may begin with one query, sample the results, and use the best samples to "re-seed" the search with more specific texts. For instance, in the case of research about property, a similarity measure was used to rank all Hansard debates as more or less similar to seed debates that were known texts about property in England, Scotland, and Ireland. The researcher sampled the results that were classified by the algorithm at different thresholds of similarity—the 1%, 5%, 10%, etc. most similar to the seed texts. From those results, the scholar collected by hand a set of exemplary texts for later analysis. This new collection of texts was both material for guided reading, and was used to "re-seed" the search process anew.

The process of continuously "checking" the work of the computer allows the expert to judge better whether and how the resulting subcorpus fits the scholarly questions at hand. Sampling the results in a structured, regular process allows the scholar to assess the results of a search confidently. Repeating the process of unsupervised matching and human sampling creates a virtuous cycle whereby the scholar gradually approaches a subcorpus ideally suited to her research question. Thus a process of critical search for digital history must fix iterative modeling, reading, and analysis of the data into the *habitude* of the modern scholar.

Results in Brief

In the case study of property in Britain, secondary sources supply a dozen possible candidates for the moment when the discourse of property changed in Britain:

[37] Eric P. S. Baumer et al., "Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence?," *Journal of the Association for Information Science and Technology* 68, no. 6 (June 1, 2017): 1397-1410.

1815, with the publication of major utilitarian pamphlets on property; 1837, with the Anti-Corn Law League; 1849, with the Encumbered Estates Act; the 1850s, with the aftermath of famine in Ireland; 1870 with the Landlord and Tenant Act for Ireland; 1881 and 1886 with the Irish Land Acts, and so on. Each history of property tends to assert that its unique clique of writers or its particular regional tradition was the most important for Britain, and there is little way for a scholar to choose between them except according to individual instinct or habit.

Digital tools promise to free the scholar from idiosyncracy of interpretation, by arming her with specific information about how these events were situated within a shifting lexicon of property. Yet, as we have seen, naïve use of the digital tools produces multiple conclusions about when and how the lexicon of property changed: perhaps because culture is multiple rather than unitary, and different parts of the lexicon were, in fact, evolving at every point in history.

Through critical search, the scholar gains some insight into which parts of the property question were evolving at which time, and thus gains confidence and insight about explaining the apparent explosion of a lexicon of property around 1880. When hand-picked titles were selected during the process of "Guided Reading," the resulting list of titles amplified a sense of a discontinuity around 1880 and allowed an even closer definition of what the relevant turning points were. The vast majority of these hand-selected titles, where parliamentary debates seemed to directly address questions of eviction, rent, and ownership, came from the period after 1881, after the Irish Land League and its "No Rent Manifesto" articulated a program of nationally-coordinated social action to force down property prices in Great Britain. There were surges of selected debates in particular years during the 1880s: in 1881 as the Land League organized; a trough in 1884, then peaking for the entire century in 1886-7 (the beginning of the "Plan of Campaign," an Irish program of joint social and parliamentary activism for land reform and the repeal of acts criminalizing public protest and the freedom of the press), dying down again in 1890 and rising almost linearly from 1890 until 1904. Examination of documents after iteration through various searches allows the researcher to assert a particular periodization with greater precision and clarity.

In the case study of a search for property in parliament, critical search produced in a very different set of material for guided reading than did either simple keyword search or topic modeling do when unaccompanied by an iterative process of seeding, winnowing, and guided reading. Naïve keyword searching tended to return, for further reading, an ungainly list of pieces of legislation related to rent, property, and agriculture, with the Irish and English legislation profiled, which is useful for establishing a chronology of the most intense debates in which rent and eviction were named, but less useful for documenting the variety of debates in which issues of rent, eviction, and property transpired. The results of naive processes reproduced a geographical bias towards England, Scotland, and Ireland, a bias inherent in the archive itself.

Unlike naïve search, the process of critical search guided the reader to a much richer caricature of the property question. The results of this process contained debates that show a marked international footprint. A debate about property that took place around the entire geography of British empire, encompassing a debate about the Egyptian sale of domain lands, and many details about police actions in Ireland at the time of the land reform, about tenancy in Bengal, Zululand farm allotments, Zanzibar land disputes, access to mountains in Scotland, the cadastral survey of Bihar, and the need for a land title registry in Ceylon, the infamous "hut tax" in central and British East Africa, the cultivation of wastelands in India, the settlement of colonists on new lands opened up after the Boer War in South Africa, Welsh colonists in Patagonia, irrigation in India, evicted ryots in Madras, the sale of public lands in the Straits settlement of Singapore, and street improvements in Kingston, Jamaica. Critical search allowed the scholar to navigate from the vast sea of agricultural and taxation documents about property to those texts that were "surprising" enough according to the algorithm to involve the plurality of ways that property was handled around emprie. This new subcorpus of texts—an "imperial property" subcorpus—can then be analysed with relative confidence of its exemplarity.

Critical search also enhances the reader's confidence in the periodization she believes to typify a conversation. The "Winnowing" of different methods allows the scholar to carefully compare the bias that different algorithms bring to questions of period. As figures 3-4 show, different divergence measures suggest different chronologies, so divergence cannot be trusted to supply a verdict about when the property question emerged and how. In this case, counting keywords underscored the sense of a historical discontinuity around 1880, which was dramatized in a keyword search of debate titles; the keyword count is relatively transparent, as a marker of a new lexicon of property, and so in the case of periodization, keyword count is preferred.

As hoped, the critical search process indeed returns parameters for an overview of social experience. It gives the advantage of a wide, contextual background to whatever close reading results at the end of the process. To define corpus, subcorpus, and research question precisely enough that scholars may be confident in the results of any models based on them is to raise the bar of knowledge.

Critical Search and Scholarly Transparency

Perhaps the most profound question raised by the use of digital tools in history is what it means to be fully transparent about our interpretive choices. In the world of social and cultural history, transparency about the scholar's bias typically took the form of a trail of footnotes to cultural anthropology or feminist theory wherein the scholar laid bare her intellectual influences, and perhaps announced an agenda for recovering the silenced voices of the past. Digital scholars too may come with announce such perspectives. But they also have the opportunity to explain how that orientation guided their maneuvers through the digital archives, caused the selection of a particular algorithms or a search for a particular lexicon, with the potential results of correcting for the biases of the past with an enhanced sensitivity that is not entirely their own.

Critical search means adopting algorithms to the research agendas we already have—feminist, subaltern, environmental, diplomatic, and so on—and searching out those tools and parameters that will enhance our prosthetic sensitivity to the multiple dimensions of the archive. Documenting the choice of seed, algorithm, cut-offs, and iteration can go a long way towards a disciplinary practice of transparency about how we understand the canon, how we develop a sensitivity to new research agendas, and how we as a field pursue the refinement of our understanding of the past.

By calling for the documentation of choices around different algorithms and their results, critical search can form the basis for a rigorous, statistically diverse overview of subject matter and time periods, making visible and transparent choices about research such as the use of secondary sources and canonical texts. In this way, digital research can build upon the findings of earlier generations, generalizing upon them or problematizing them at enormous scale.

Critical search promises transparency in these findings, making good on the commitment of earlier generations of scholars to replicability in humanistic research and even radically extending that commitment to the every-day choices made by scholars. Traditionally, the scholar plucks events, characters, and research questions out of the archive by a combination of individual proclivity, expert guidance, happy accident, and close reading; a good research project is one where a wide enough variety of sources materialize to make thick reading possible. As this article demonstrates, the virtues of iterative rigor, broad contextual reading, and curious questioning as the marks of inspiring scholarship, and these will remain important qualifiers of individual talent in a digital age.

In explaining how each visualization is the result of particular choices in the accumulation of starting point, keywords, secondary sources, algorithms, and their deployment, critical search will tend to move the reading and interpretation of data-driven visualizations away from a naïve reading, where visualizations appear to propose the view from nowhere, performing what Donna Harraway dubbed "the god trick."[38] Instead, some visualizations may be used to compare the bias of different measurements (Figure 4). Others will to explore different dimensions of a corpus rendered visible by algorithms, but visualizations juxtaposed will illuminate how the same question can be asked different ways (Tables 1-3, 4).

In calling for transparent documentation of the choices that go into research, however, critical search thus does not propose to eliminate the scholar's personal biases or to render all historical research transcendentally objective. The scholar, after all, chooses the seed texts, algorithms, and their cut-offs; and it is only the scholar who chooses and reads the new texts supplied by this process. Subjectivity and opportunities of individual insight remain at every level.

[38] Donna Haraway, ”Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective," *Feminist Studies* 14, no. 3 (1988): 581.